
Machine Learning

Topic 4: Linear Regression Models

(with ideas and a few images from wikipedia and books by Alpaydin, Duda/Hart/Stork, and Bishop)

General Regression Learning Task

There is a set of possible examples $X = \{\vec{x}_1, \dots, \vec{x}_n\}$

Each example is an k -tuple of attribute values

$$\vec{x}_1 = \langle a_1, \dots, a_k \rangle$$

There is a target function that maps X onto some **real value** Y

$$f : X \rightarrow Y$$

The DATA is a set of tuples $\langle \text{example}, \text{target function values} \rangle$

$$D = \{ \langle \vec{x}_1, f(\vec{x}_1) \rangle, \dots, \langle \vec{x}_m, f(\vec{x}_m) \rangle \}$$

Find a **hypothesis** h such that...

$$\forall \vec{x}, h(\vec{x}) \approx f(\vec{x})$$

Reminder: Expected Value

The expected value of a random variable X , $\mathbb{E}[X]$, is the mean of the probability distribution of X .

For a discrete random variable, where $p(x)$ is the probability of x , and the sum is over all possible x :

$$\mathbb{E}[X] = \sum_x xp(x)$$

For a continuous random variable, where $f(x)$ is the probability density function of X :

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Why use a linear regression model?

- Easily understood
- Interpretable
- Well studied by statisticians
 - many variations and diagnostic measures
- Computationally efficient

Linear Regression Model

Assumptions: The observed response (dependent) variable, r , is the true function, $f(x)$, with additive Gaussian noise, ε

Observed response $r = f(\vec{x}) + \varepsilon$

Where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

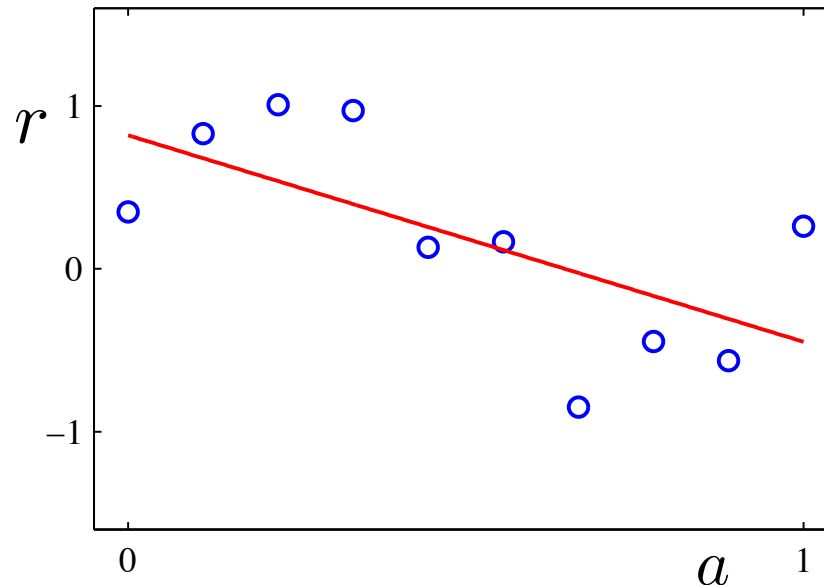
Reminder: this is a “Normal”
(Gaussian) distribution

Simple Linear Regression

- x has 1 attribute (predictor variable)
- Hypothesis function is a line:

$$h(\vec{x} | w_1, w_0) = w_0 + w_1 a$$

Example:

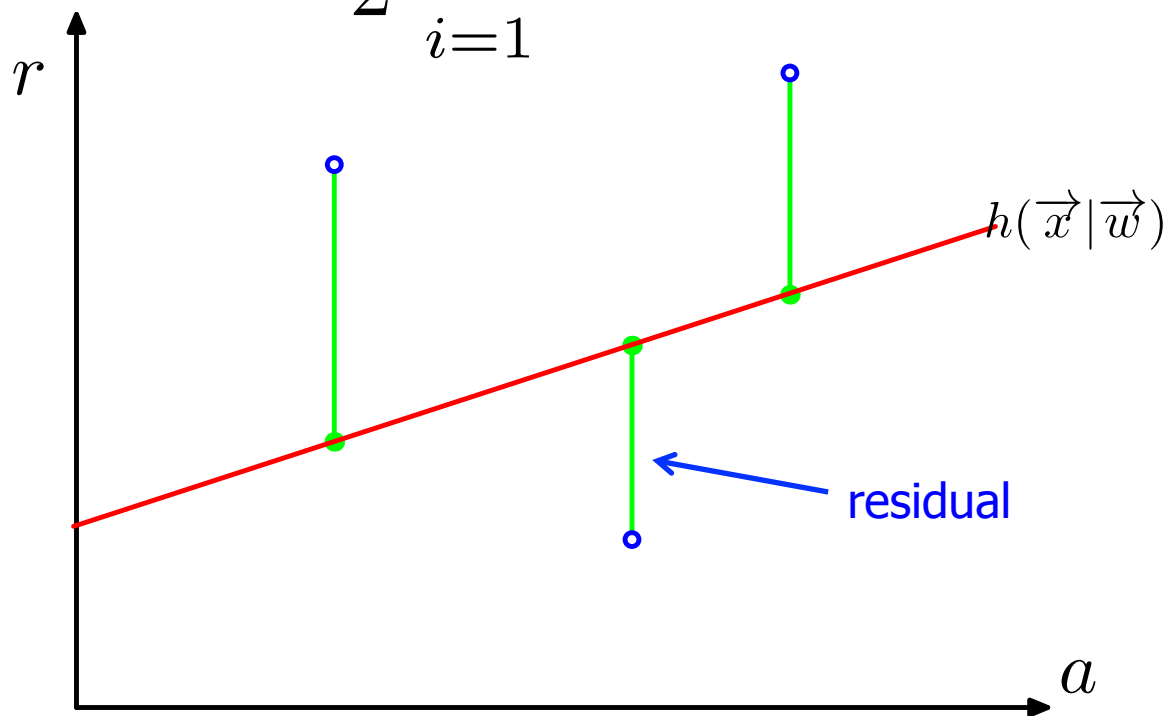


Simple Linear Regression

Typically estimate parameters by minimizing sum of squared residuals (a.k.a. least squares):

$$SSE = \frac{1}{2} \sum_{i=1}^m [r_i - h(\vec{x}_i | \vec{w})]^2$$

← number of training examples



Linear Regression Model

Assumption: The expected value of the response variable, $\mathbb{E}[R|\vec{x}]$, is approximately a linear combination of independent variables (the attributes/features).

Therefore our, hypothesis function in linear regression is:

$$\begin{aligned} h(\vec{x}|\vec{w}) &= \mathbb{E}[r|\vec{x}] \\ &= w_0 + w_1 a_1 + w_2 a_2 + \cdots + w_k a_k \end{aligned}$$

$$p(r|\vec{x}) \sim \mathcal{N}(h(\vec{x}|\theta), \sigma^2)$$

Linear Regression Model

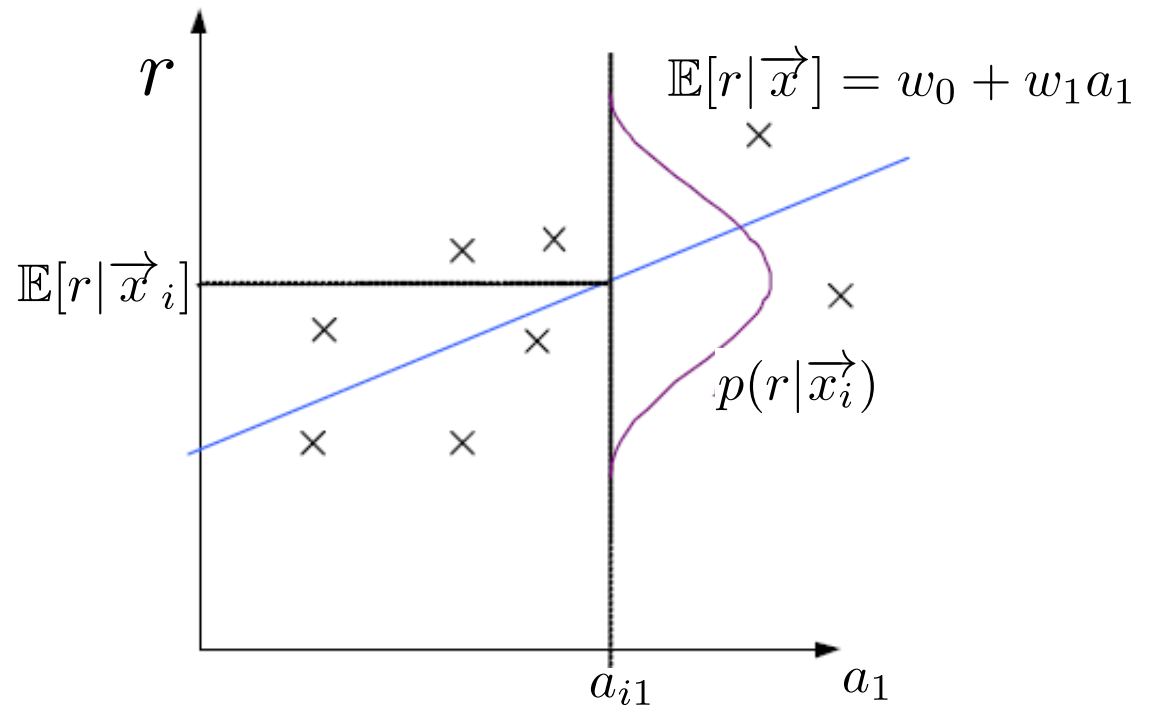
observed response variable: $r = f(\vec{x}) + \varepsilon$

noise: $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

estimator: $h(\vec{x} | \vec{w}) = \mathbb{E}[r | \vec{x}] = w_0 + w_1 a_1 + w_2 a_2 + \dots + w_k a_k$

distribution of response: $p(r | \vec{x}) \sim \mathcal{N}(h(\vec{x} | \theta), \sigma^2)$

Example with 1
attribute/predictor variable:



Simple (Univariate) Linear Regression

Take derivatives, set to 0, and solve:

Squared residuals:
$$SSE = \frac{1}{2} \sum_{i=1}^m [r_i - h(\vec{x}_i | \vec{w})]^2$$

Derivatives with respect to w :

$$\frac{\partial SSE}{\partial w_0} = \sum_i r_i = mw_0 + w_1 \sum_i a_i$$
$$\frac{\partial SSE}{\partial w_1} = \sum_i r_i a_i = w_0 \sum_i a_i + w_1 \sum_i a_i^2$$

Analytic solution:

$$\mathbf{A} = \begin{bmatrix} m & \sum_i a_i \\ \sum_i a_i & \sum_i a_i^2 \end{bmatrix} \quad \vec{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \vec{y} = \begin{bmatrix} \sum_i r_i \\ \sum_i r_i a_i \end{bmatrix}$$

$$\vec{w} = \mathbf{A}^{-1} \vec{y}$$

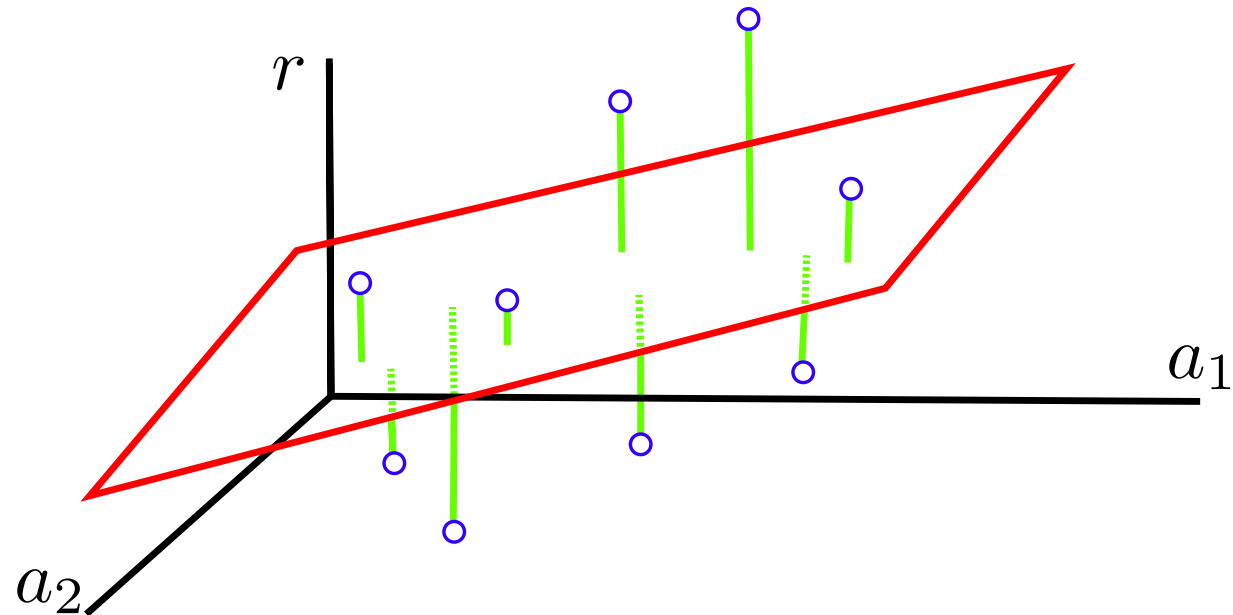
Multiple (Multivariate*) Linear Regression

*NOTE: In statistical literature, multivariate linear regression is regression with multiple outputs, and the case of multiple input variables is simply “multiple linear regression”

- Many attributes
- Response is a hyperplane

$$h(\vec{x} | w_0, w_1, \dots, w_k) = w_0 + w_1 a_1 + w_2 a_2 + \dots + w_k a_k$$

Example with
2 attributes:



Multiple (Multivariate) Linear Regression

Parameter estimation (analytically minimizing sum of squared residuals):

One training example

$$\mathbf{X} = \begin{bmatrix} 1 & a_{1,1} & a_{1,2} & \dots & a_{1,k} \\ 1 & a_{2,1} & a_{2,2} & \dots & a_{2,k} \\ \vdots & & & & \\ 1 & a_{m,1} & a_{m,2} & \dots & a_{m,k} \end{bmatrix} \quad \vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}$$

$$\vec{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \vec{r} \quad \leftarrow \text{“normal equations”}$$

Linear Regression Model

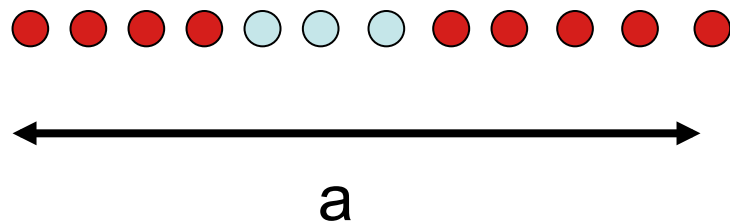
The attributes may be of higher order, as long as the coefficients, \vec{w} , are not. E.g, polynomial regression is still a linear regression model:

$$h(\vec{x}|w_k, \dots, w_2, w_1, w_0) = w_0 + w_1 a + w_2 a^2 + \dots + w_k a^k$$

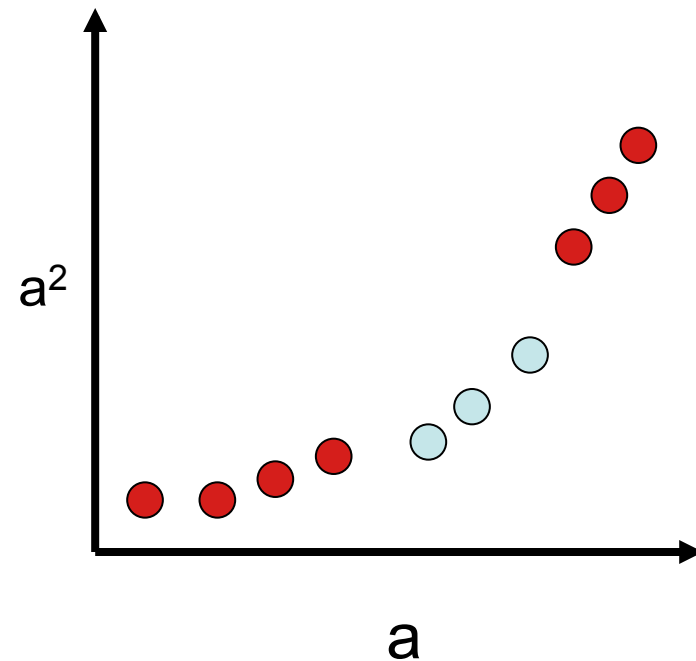
(How can it be linear with exponents in the formula?....let's see)

Polynomial Expansion

You can project a single one-dimensional predictor variable into a higher dimensional space by creating new dimensions which are simply power functions of the single predictor variable.



Mapping a onto $\{a^2, a\}$



Polynomial Regression

If we think of these new dimensions as new variables, then the function is still linear, and we can perform multiple linear regression in this higher dimensional space.

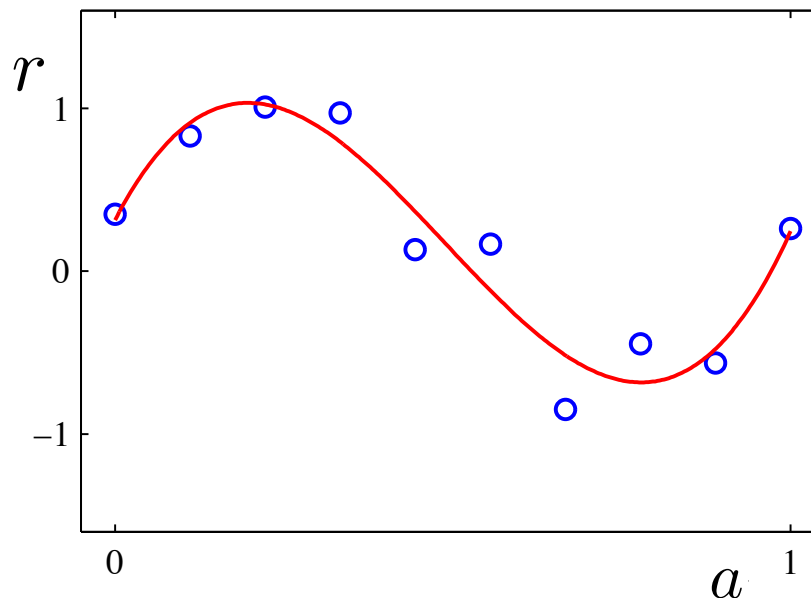
$$h(\vec{x}|w_k, \dots, w_2, w_1, w_0) = w_0 + w_1 a + w_2 a^2 + \dots + w_k a^k$$

However, when interpreted as a function of the single predictor variable, this is a non-linear function that is a linear combination of polynomial basis functions (i.e. it is still linear regression).

Polynomial Regression

- Model the relationship between the response variable and the attributes/predictor variables as a k^{th} -order polynomial. While this can model non-linear functions, it is still linear with respect to the coefficients.

$$h(\vec{x}|w_k, \dots, w_2, w_1, w_0) = w_0 + w_1 a + w_2 a^2 + \dots + w_k a^k$$



Polynomial Regression

Parameter estimation (analytically minimizing sum of squared residuals):

One training example

$$\mathbf{D} = \begin{bmatrix} 1 & a_1 & a_1^2 & \dots & a_1^k \\ 1 & a_2 & a_2^2 & \dots & a_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_m & a_m^2 & \dots & a_m^k \end{bmatrix} \quad \vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}$$

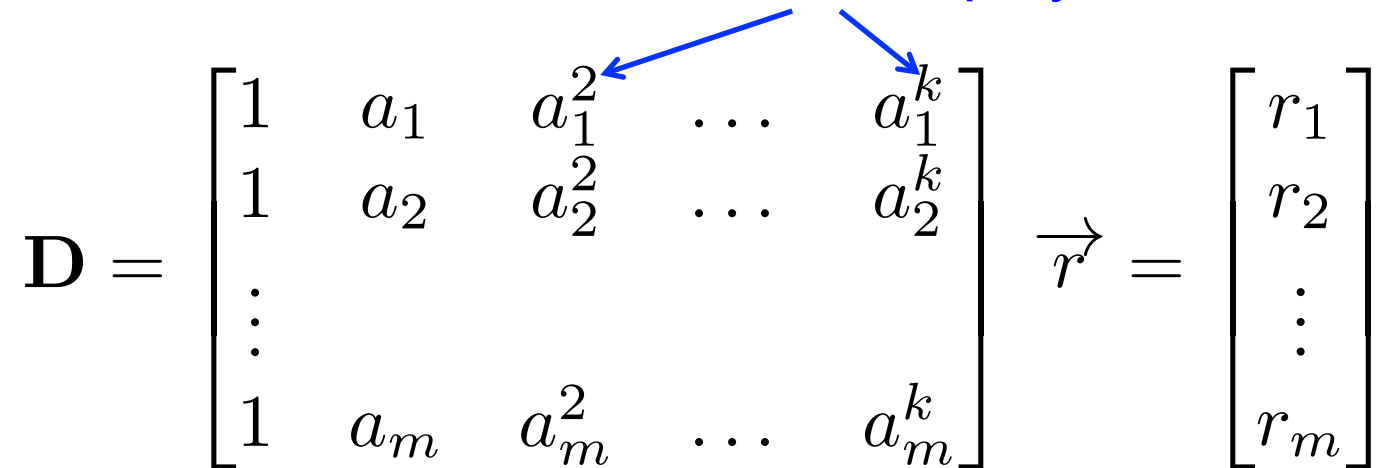
(Note, there is only 1 attribute a for each training example.
Those superscripts are powers, since we're doing polynomial regression)

$$\vec{w} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \vec{r} \quad \leftarrow \text{“normal equations”}$$

Polynomial Regression

Parameter estimation (analytically minimizing sum of squared residuals):

What makes it polynomial

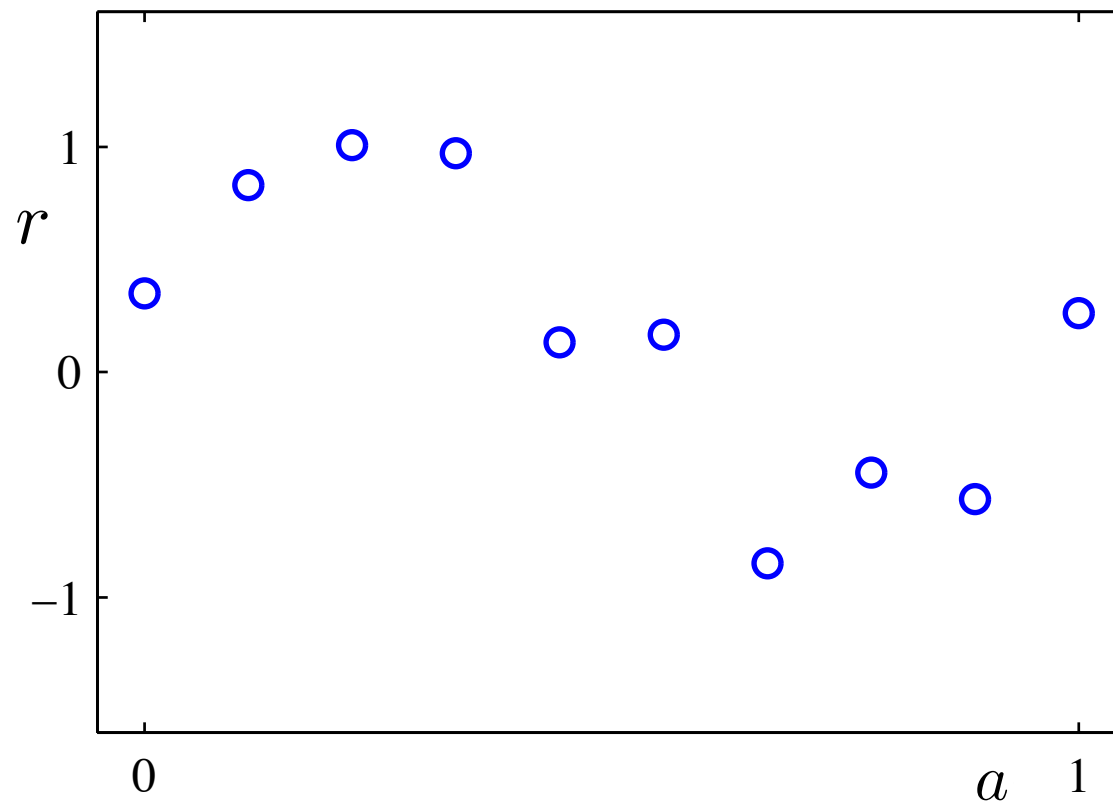

$$\mathbf{D} = \begin{bmatrix} 1 & a_1 & a_1^2 & \dots & a_1^k \\ 1 & a_2 & a_2^2 & \dots & a_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & a_m & a_m^2 & \dots & a_m^k \end{bmatrix} \quad \vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}$$

(Note, there is only 1 attribute a for each training example.
Those superscripts are powers, since we're doing polynomial regression)

$$\vec{w} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T \vec{r} \quad \leftarrow \text{"normal equations"}$$

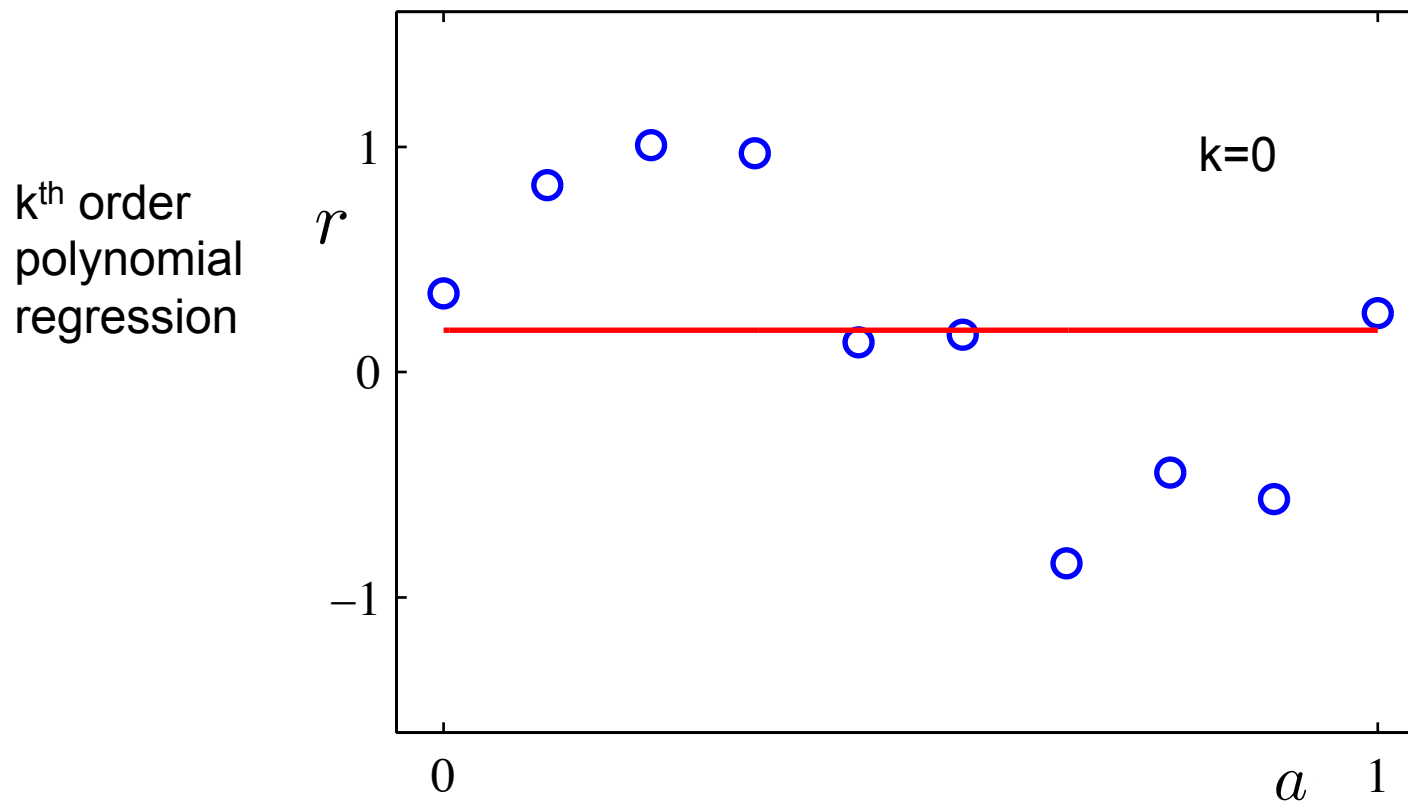
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



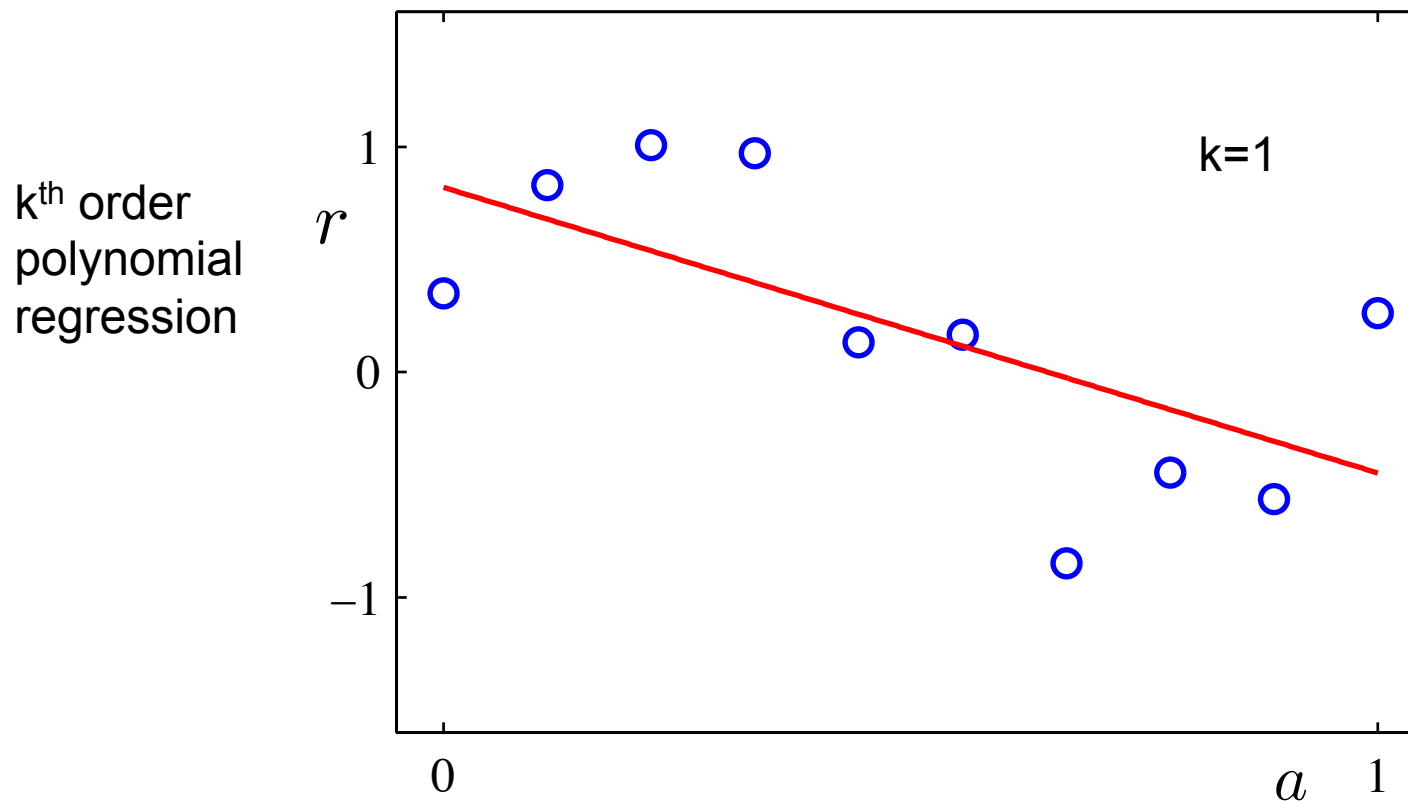
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



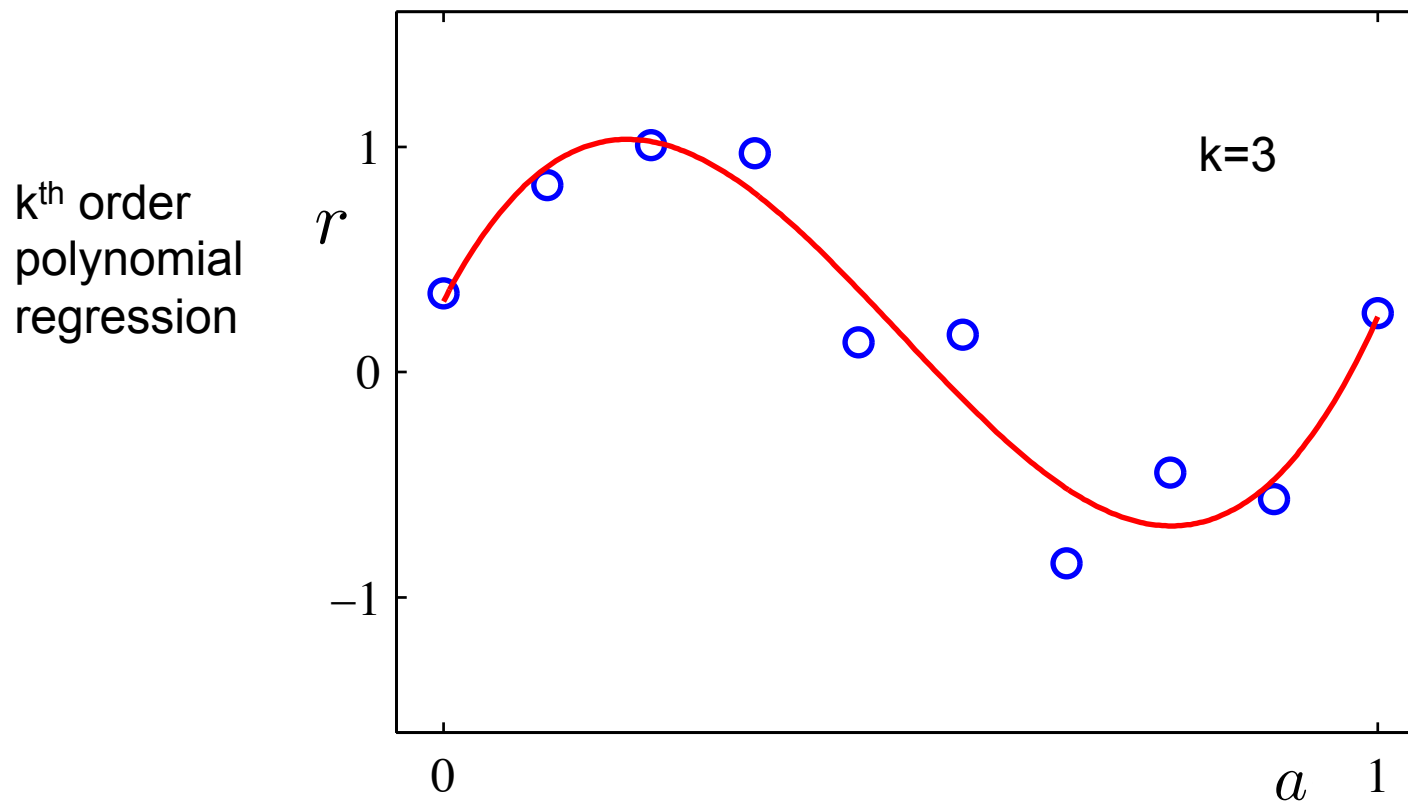
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



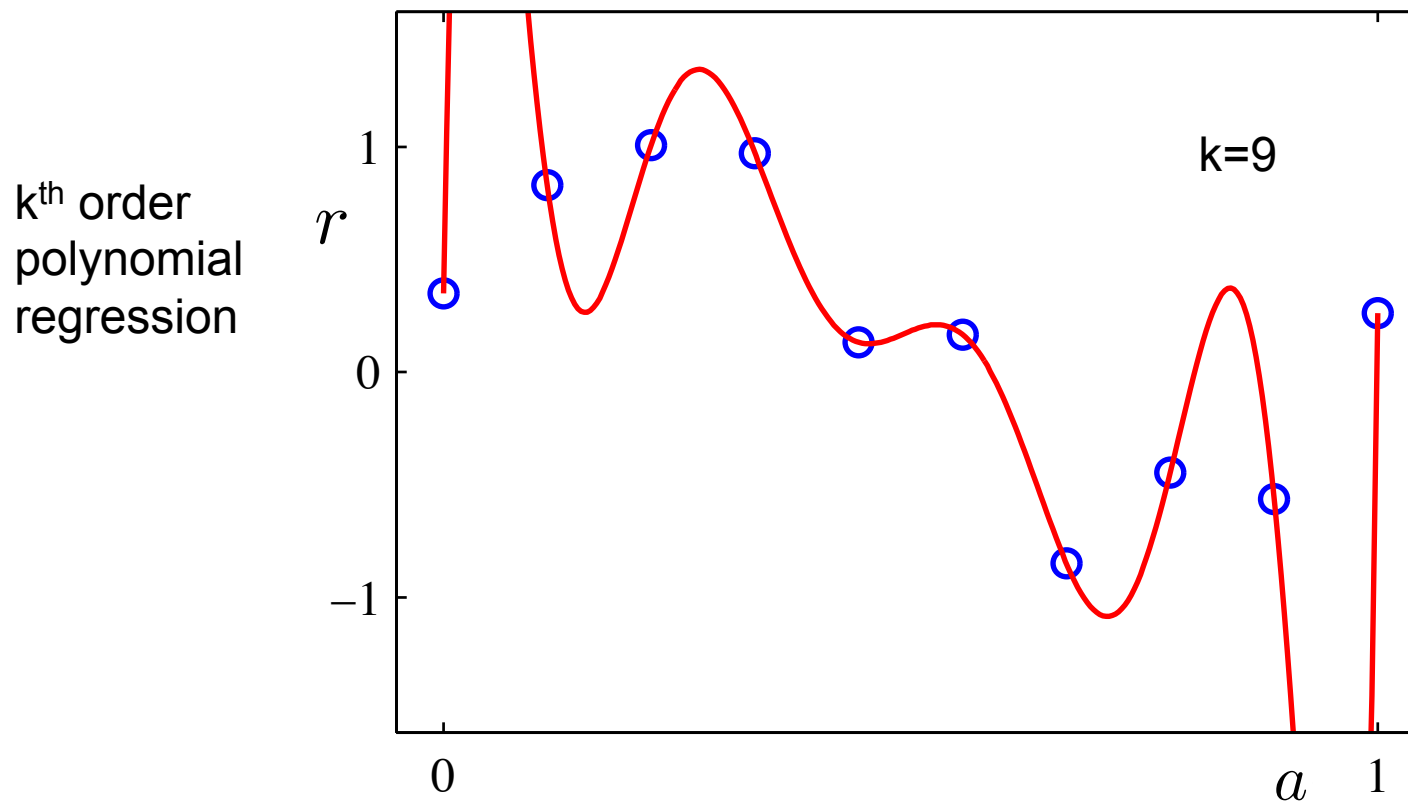
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



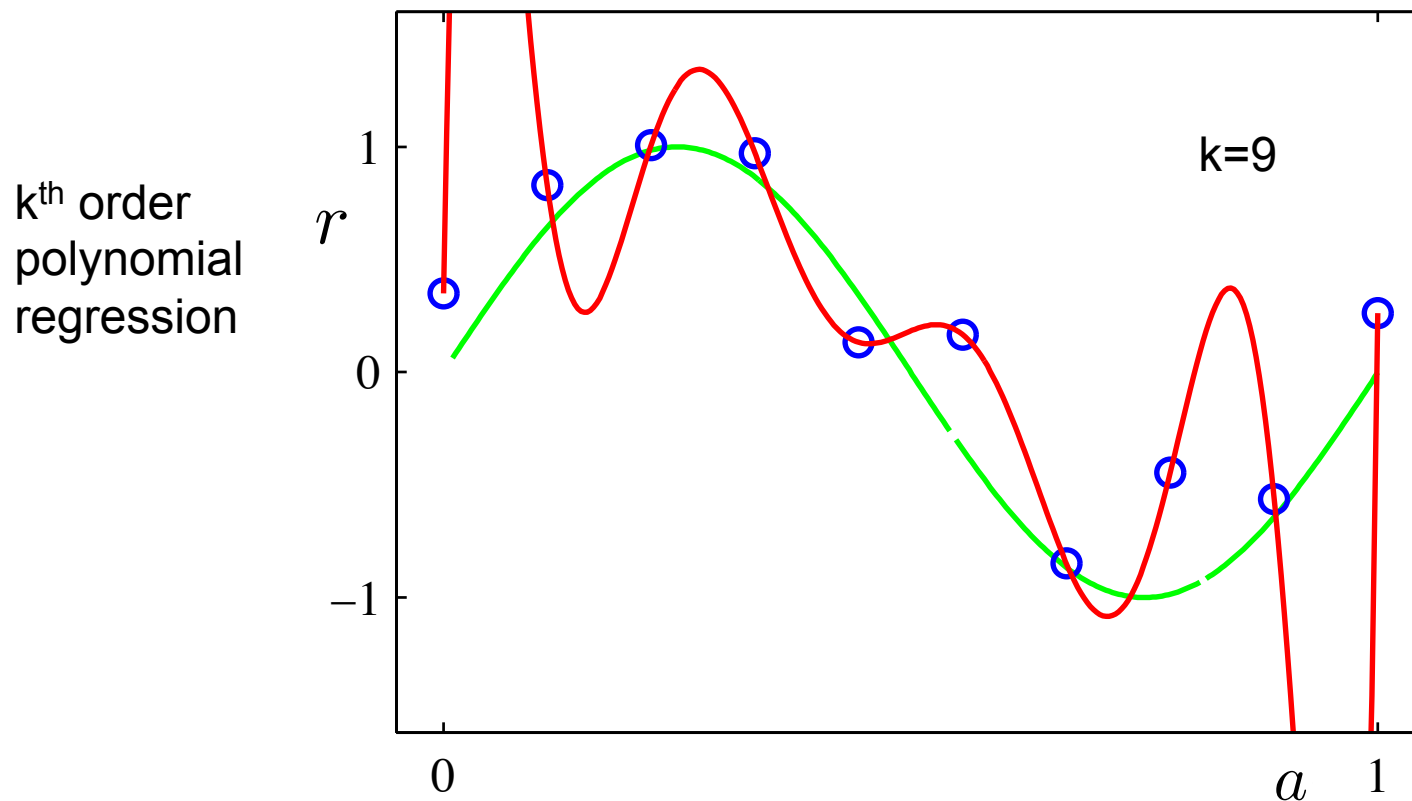
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



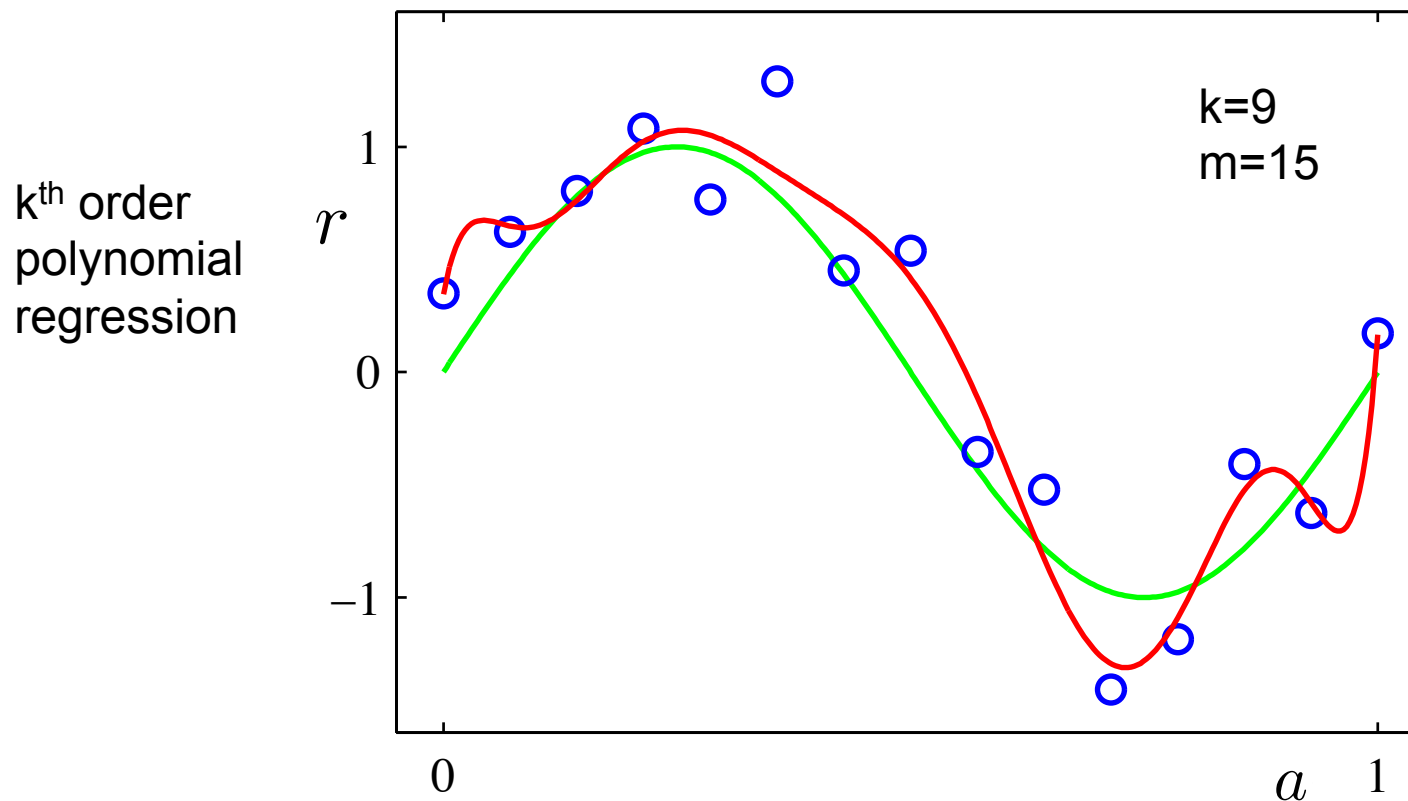
Tuning Model Complexity: Example

What is your hypothesis for $f(x)$?



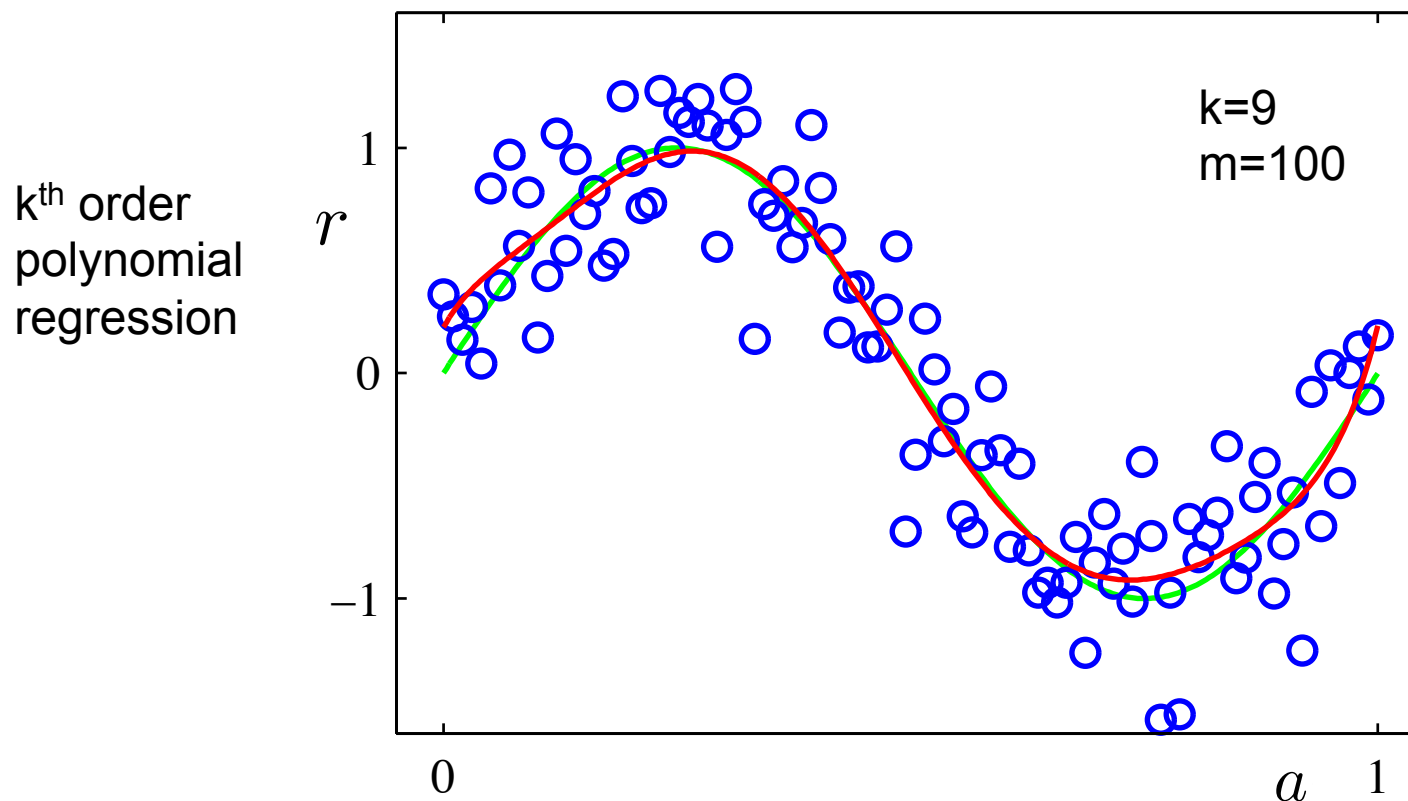
Tuning Model Complexity: Example

What happens if we fit to more data?



Tuning Model Complexity: Example

What happens if we fit to more data?



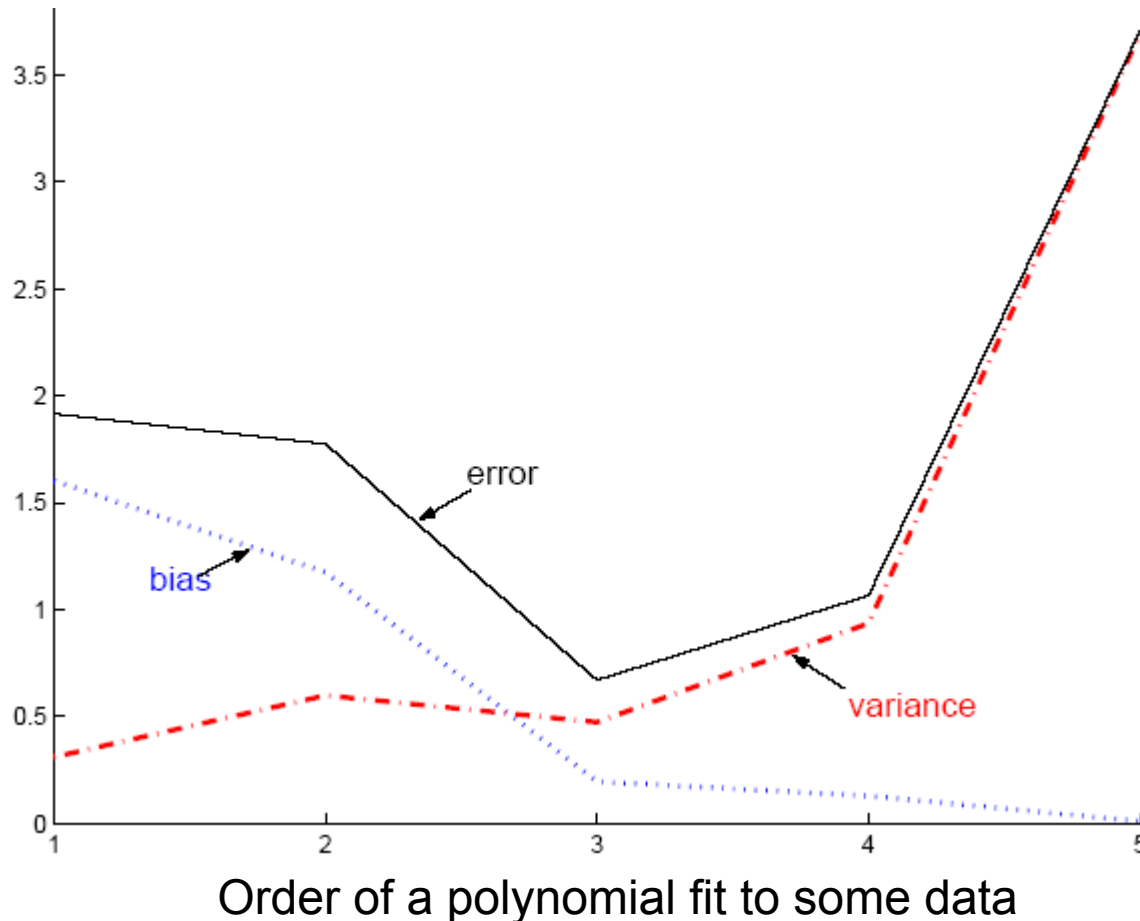
Bias and Variance of an Estimator

- Let X be a sample from a population specified by a true parameter θ
- Let $d=d(X)$ be an estimator for θ

$$\mathbb{E}[(d - \theta)^2] = \mathbb{E}[(d - \mathbb{E}[d])^2] + (\mathbb{E}[d] - \theta)^2$$

mean square error *variance* *bias*²

Bias and Variance



As we **increase complexity**, **bias decreases** (a better fit to data) and **variance increases** (fit varies more with data)

Bias and Variance of Hypothesis Fn

- Bias:

Measures how much $h(x)$ is wrong disregarding the effect of varying samples (This is the statistical bias of an estimator. This is NOT the same as inductive bias, which is the set of assumptions that your learner is making)

- Variance:

Measures how much $h(x)$ fluctuate around the expected value as the sample varies.

NOTE: These concepts are general machine learning concepts, not specific to linear regression.

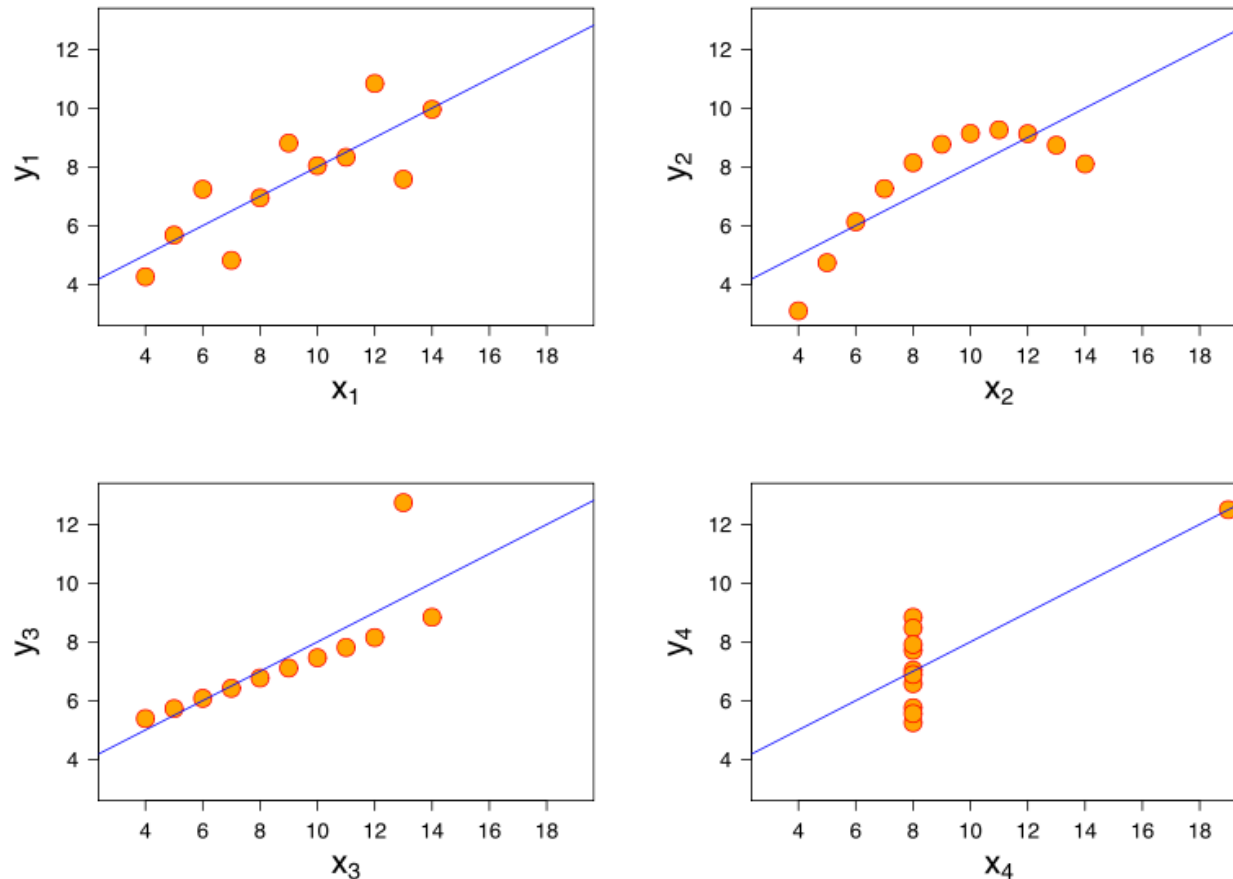
Coefficient of Determination

- the **coefficient of determination**, or **R^2** indicates how well data points fit a line or curve. We'd like R^2 to be close to 1

$$E_{RSE} = \frac{\sum_{i=1}^m [r_i - h(\vec{x}_i | \vec{w})]^2}{\sum_{i=1}^m [r_i - \bar{r}]^2} \quad \leftarrow \text{relative square error}$$

$$R^2 = 1 - E_{RSE} \quad \leftarrow \text{Coefficient of determination (r-square)}$$

Don't just rely on numbers, visualize!



For all 4 sets: same mean and variance for x , same mean and variance (almost) for y , and same regression line and correlation between x and y (and therefore same R-squared).

Summary of Linear Regression Models

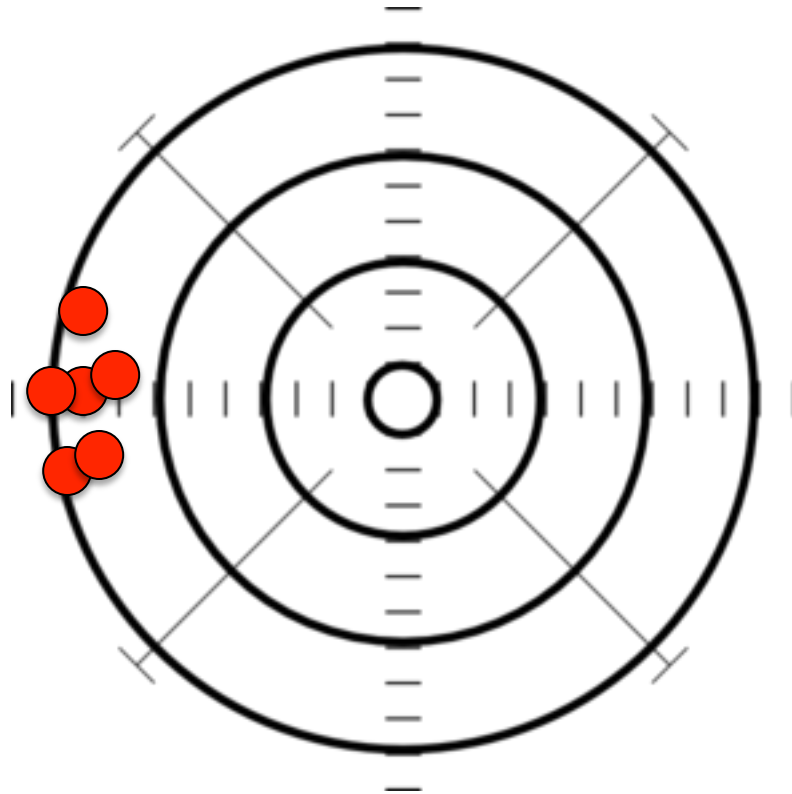
- Easily understood
- Interpretable
- Well studied by statisticians
- Computationally efficient
- Can handle non-linear situations if formulated properly
- Bias/variance tradeoff (occurs in all machine learning)
- Visualize!!
- GLMs

Appendix

(Stuff I couldn't cover in class)

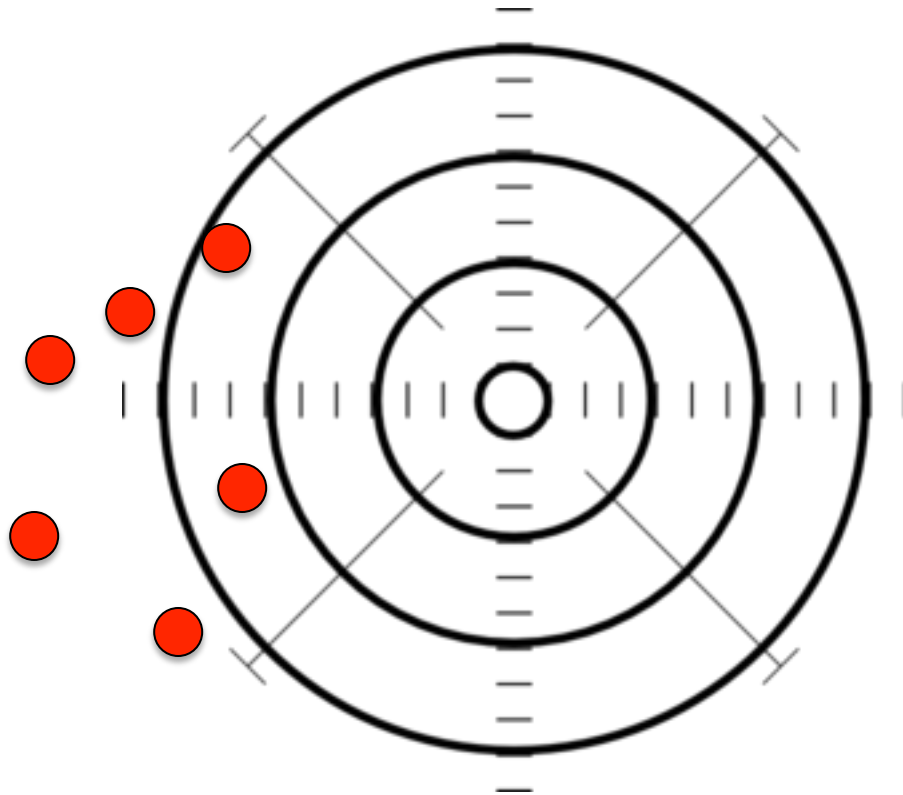
Bias and Variance

high bias, low variance



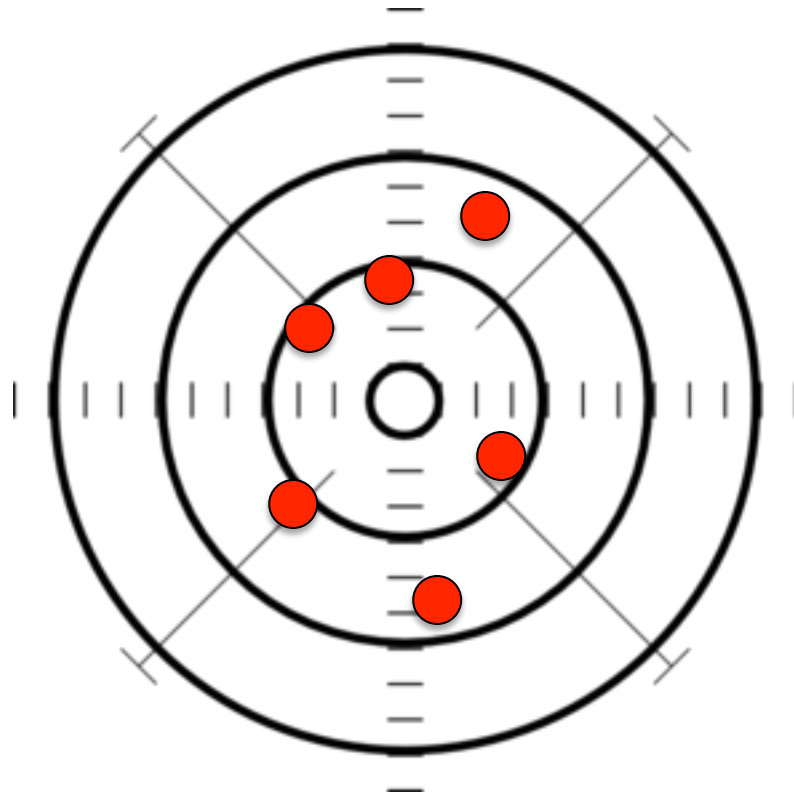
Bias and Variance

high bias, high variance



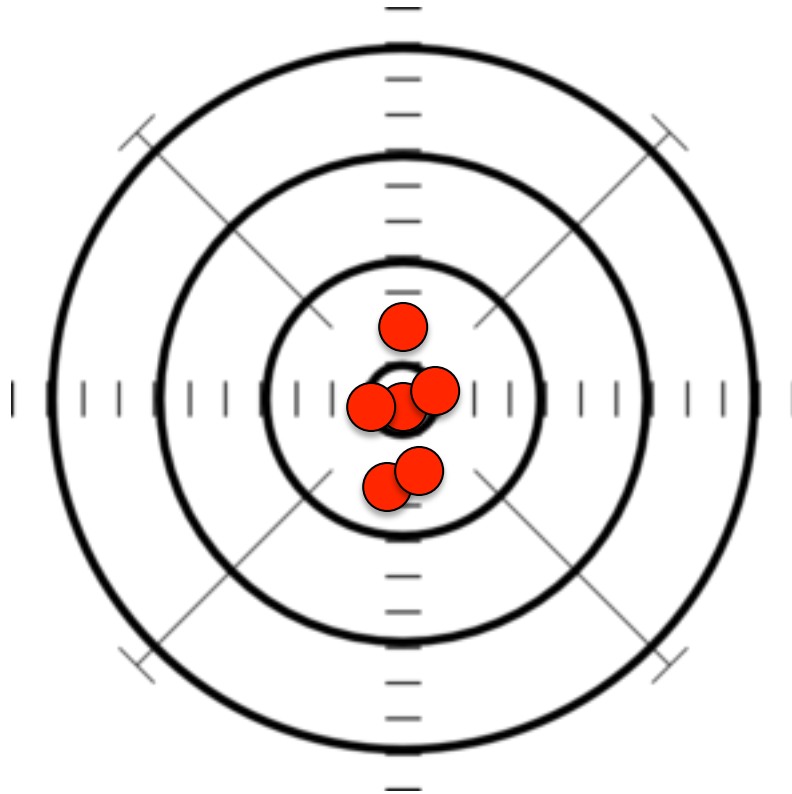
Bias and Variance

low bias, high variance



Bias and Variance

low bias, low variance



Bias and Variance in Linear Regression

$$\mathbb{E} \left[(r - h(\vec{x}))^2 | \vec{x} \right] = \mathbb{E} \left[(r - \mathbb{E}[r | \vec{x}])^2 | \vec{x} \right] + (\mathbb{E}[r | \vec{x}] - h(\vec{x}))^2$$

mean square error between response and hypothesis *variance of noise*

squared error between "ideal" predictor and hypothesis*

$$\mathbb{E}_{\mathcal{X}} \left[(\mathbb{E}[r | x] - h(x))^2 | x \right] = (\mathbb{E}[r | x] - \mathbb{E}_{\mathcal{X}}[h(x)])^2 + \mathbb{E}_{\mathcal{X}} \left[(h(x) - \mathbb{E}_{\mathcal{X}}[h(x)])^2 \right]$$

mean square error between "ideal" predictor and hypothesis (averaged over all possible datasets)* *bias²*

variance

*"ideal" based on our assumption that the noise is 0 mean.

NOTE: If we knew the true function then we can replace $\mathbb{E}[r|x]$ for it in the second equation

Bias and Variance

- Bias:

Measures how much $h(x)$ is wrong disregarding the effect of varying samples

high bias  underfitting

- Variance:

Measures how much $h(x)$ fluctuate around the expected value as the sample varies.

high variance  overfitting

There's a trade-off between bias and variance

Ways to Avoid Overfitting

- Simpler model
 - E.g. fewer parameters
- Regularization
 - penalize for complexity in objective function
- Fewer features
- Dimensionality reduction of features (e.g. PCA)
- More data...

Model Selection

- **Cross-validation:** Measure generalization accuracy by testing on data unused during training
- **Regularization:** Penalize complex models
 $E' = \text{error on data} + \lambda \text{ model complexity}$
Akaike's information criterion (AIC), Bayesian information criterion (BIC)
- **Minimum description length (MDL):** Kolmogorov complexity, shortest description of data
- **Structural risk minimization (SRM)**

Generalized Linear Models

- Models shown have assumed that the response variable follows a Gaussian distribution around the mean
- Can be generalized to response variables that take on *any* exponential family distribution (Generalized Linear Models - GLMs)